

# PLFS UPDATE 2011

Adam Manzanares, et. al.

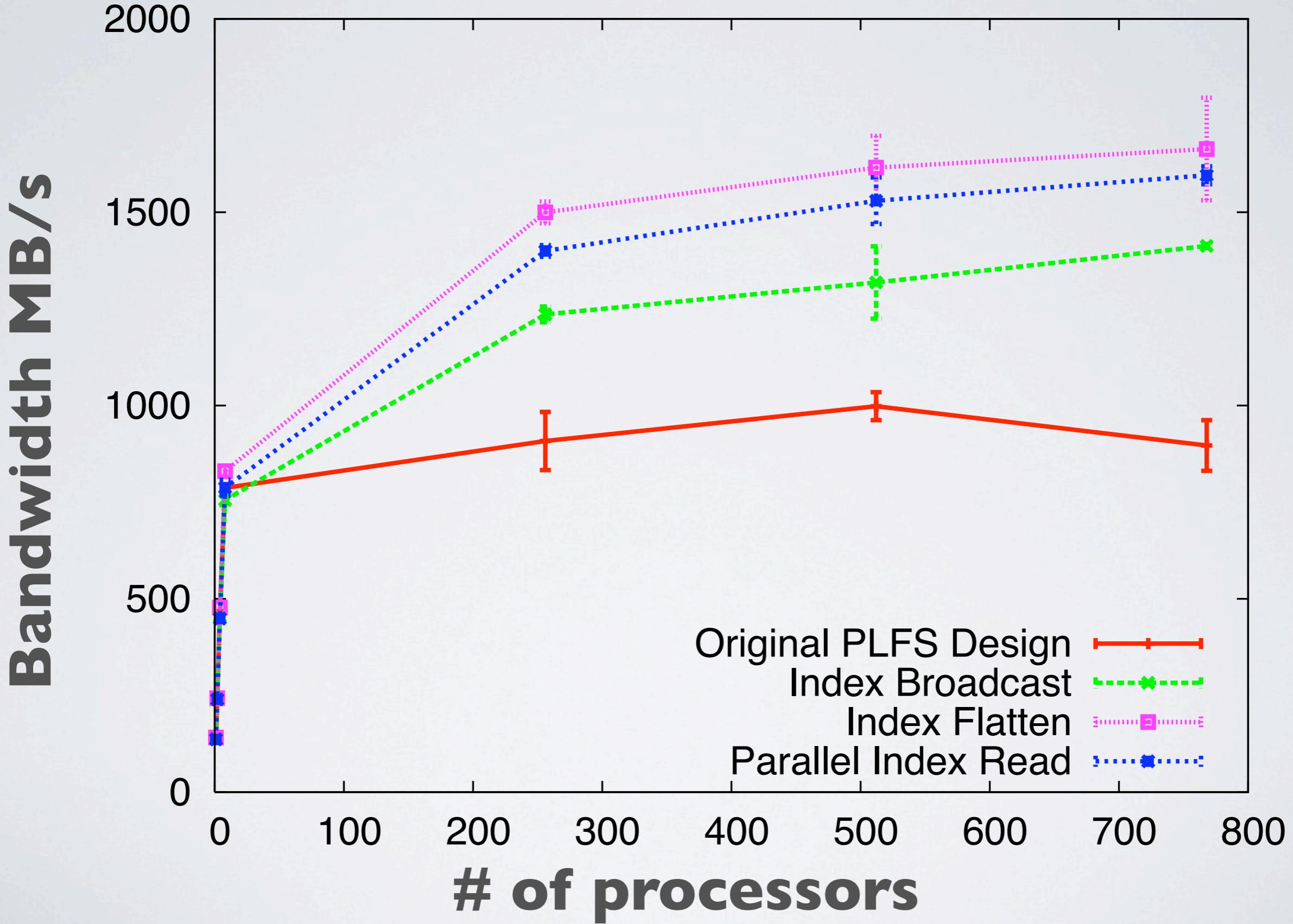
# PLFS

- Parallel Log Structured File System
  - Improves Shared File Checkpointing Bandwidth
  - Decouples writers
  - Up to 150x improvement in write bandwidth
  - Distributes metadata load
  - In many cases no read performance penalty
  - Now accessible both through FUSE and through MPI-IO (plfs adio)

# READ OPTIMIZATIONS

- Archiving
  - N processors write the file, a handful of archive servers read it
  - Increase parallelism with multi-threaded reads
    - Helps time overhead but not space ...
- Checkpoint restart
  - $N^2$  Open Problem
  - Leverage MPI to coordinate read access

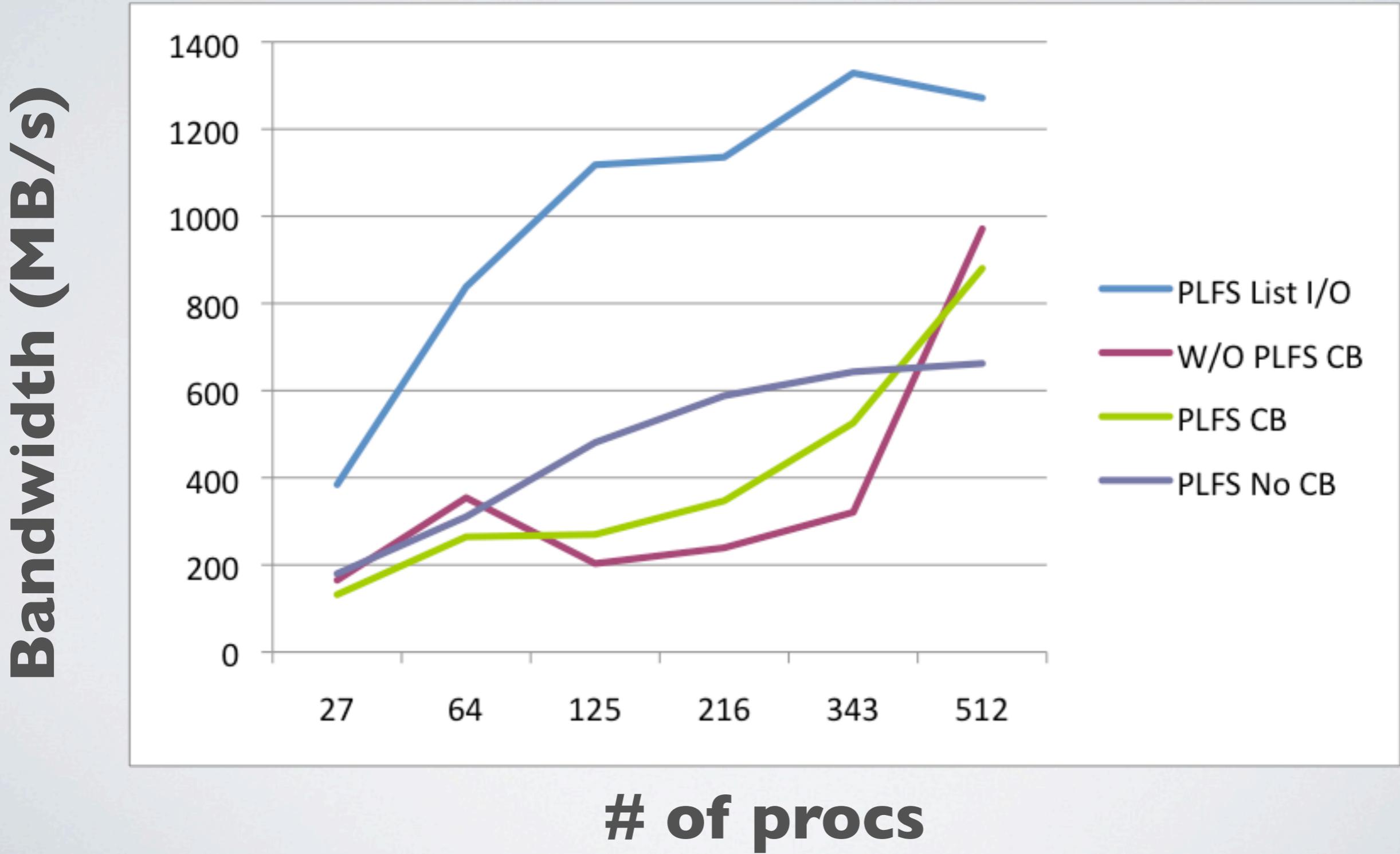
# Read BW

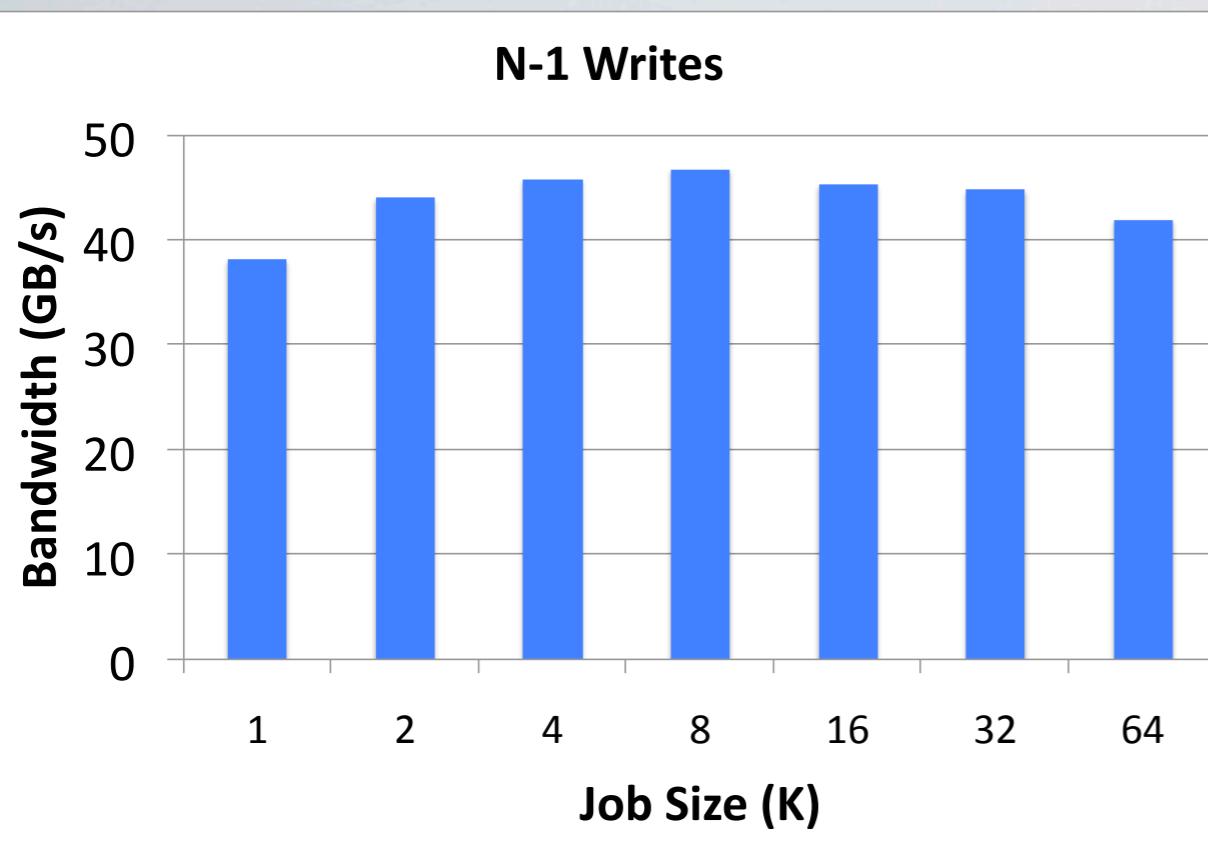


# LIST I/O

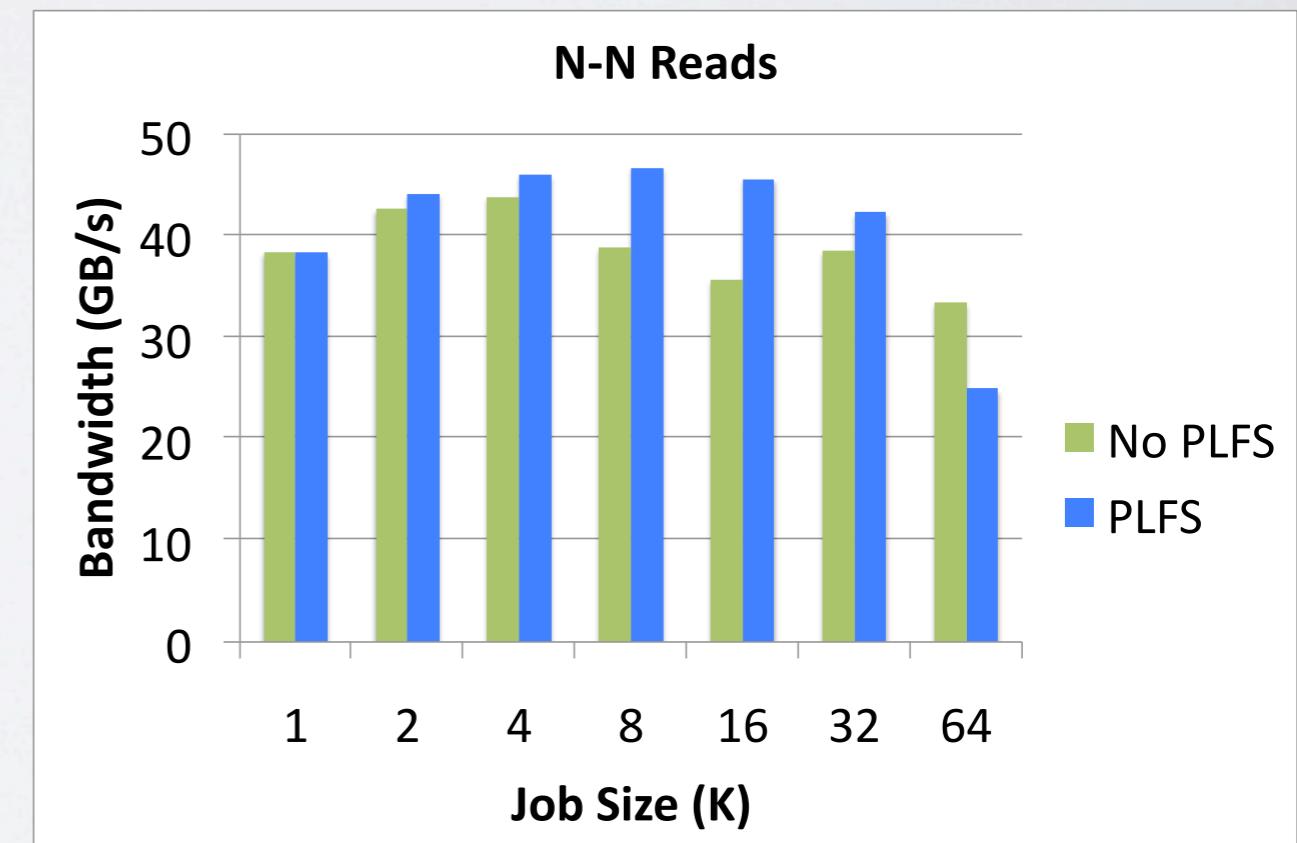
- Implemented strided writing for PLFS
  - Log structure beneficial.
- Read still a problem for very large numbers of writers (>16K)
  - Index is still too large
  - Investigating alternative representations for index
    - Capture patterns from MPI Set View

# LIST I/O WRITE BANDWIDTH





PLFS on LANL's Cielo with Ten Metadata Servers



# CURRENT STATUS

- PLFS being pushed into production on LANL's Cielo
  - Required for both N-I and N-N
- MPI-IO interface since Version 1.2
- Metadata distribution since Version 2.0